Finding and Archiving the Internet Footprint*

Simson Garfinkel† and David Cox
Naval Postgraduate School
Monterey, CA, USA

February 10, 2009

Abstract

With the move to "cloud" computing, archivists face the increasingly difficult task of finding and preserving the works of an originator so that they may be readily used by future historians. This paper explores the range of information that an originator may have left on computers "out there on the Internet," including works that are publicly identified with the originator; information that may have been stored using a pseudonym; anonymous blog postings; and private information stored on web-based services like Yahoo Calendar and Google Docs. Approaches are given for finding the content, including interviews, forensic analysis of the originator's computer equipment, and social network analysis. We conclude with a brief discussion of legal and ethical issues.

Keywords: Forensics, Search, Historical Record, Information Gathering

1 Introduction

With the introduction of home computers and electronic typewriters in the late 1970s, archivists were forced to confront the fact that a person's "papers" might, in fact, no longer be on paper[29]. The power of word processing made writers among the first to embrace information technology outside of government and the financial sector. And because writers often made small purchases and were not constrained by prior investment, they frequently purchased equipment from small niche manufacturers whose technology did not become dominant.

As a result, preserving and cataloging the earliest electronic records consisted of two intertwined problems: the task of *finding* and copying the data off magnetic media before the media deteriorates, and the challenging of reading older and sometimes obscure formats that are no longer in widespread use[1].

Archivists are now on the brink of a far more disruptive change than the transition from paper to electronic media: the transition from personal to "cloud computing." In the very near future an archivist might enter the office of a deceased writer and find *no electronic files of personal significance*: the author's appointment calendar might split between her organization's Microsoft Exchange server and Yahoo Calendar; her unfinished and unpublished documents stored on Google Docs; her diary stored at the online LiveJournal service; correspondence archived on the Facebook "walls" of her close friends; and her most revealing, insightful and critical comments scattered as anonymous and pseudonymous comments on the blogs of her friends, collaborators, and rivals.

Although there are numerous public and commercial projects underway to find and preserve public web-based content, these projects will not be useful to future historians if there is no way to readily find the information that is of interest. And of course, none of the archiving projects are able to archive content that is private or otherwise restricted—as will increasingly be the case of personal information that is stored in the "cloud."

1.1 Outline of this paper

This paper introduces and explores the problem of finding and archiving person's *Internet footprint*. In Section 2 we define the term *Internet footprint* and provide numerous examples of the footprint's extent. In Section 3 we

*Invited paper, presented at the First Digital Lives Research Conference: Personal Digital Archives for the 21st Century, London, England, 9–11 February 2009

†Corresponding Author: slgarfin@nps.edu

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
10 FEB 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Finding and Archiving the Internet Footprint	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

With the move to ?cloud? computing, archivists face the increasingly difficult task of finding and preserving the works of an originator so that they may be readily used by future historians. This paper explores the range of information that an originator may have left on computers ?out there on the Internet,? including works that are publicly identified with the originator; information that may have been stored using a pseudonym; anonymous blog postings and private information stored on web-based services like Yahoo Calendar and Google Docs. Approaches are given for finding the content, including interviews forensic analysis of the originator?s computer equipment and social network analysis. We conclude with a brief discussion of legal and ethical issues.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**14**	

present a variety of approaches for finding the footprint. In Section 4 we discuss technical concerns for archiving the footprint.

1.2 Related Work

Web archiving has received significant exploration in recent years, including the use of proxies to collect data[42], the need for proper record management[41], and the difficulty of reconstructing lost websites from the web infrastructure[36]. Researchers have also characterizied the Web's "decay"[7]. Jatowt *et al.* have developed techniques for automatically detecting the age of a web page[28].

Juola provides a review of current authorship determination techniques[30].

There are numerous open source and commercially available face recognition products, including FaceIt by Visionics, FavesVACS by Plettac, and ImageWare Software. Zhao *et al.* [50] and Datta *et al.* [15] have both published comprehensive surveys of current research and technology.

Viégas *et al.* examined cooperation and conflict between authors by analyzing Wikipedia logs[48]. Other relevant work on Wikipedia includes analysis of participation[9] and statistical models that can predict future administrators[11].

2 The Internet Footprint

Consider the staggering range of Internet services that a person uses during the course of a year. Some of these are public publication services like BBC or CNN News—services that are little more than traditional television, radio or newspaper repurposed to the Internet, and that most Internet users access anonymously. Other services are public and highly personalized—blogs and home pages, for example. Still other services are private and personal, like an online calendar or diary. These services can be operated by an organization for its employees, such as a company running a Microsoft Exchange server, or they can be operated on a global scale for millions of users, such as Google Calendar[23].

This section considers the wide range of information that an originator may create in other computers on the Internet through their own actions—the originator's *Internet Footprint*.

2.1 The Public Identified Footprint

A person's public identified footprint is any information that they created which is online, widely available, and specifically linked to author's real name.

For originators that are authors, their public footprint almost certainly includes articles that have been published under the originator's own name in web-only publications such as *Slate Magazine*[5] or Salon.com[4]. The public footprint may also include letters to the editor. (John Updike once wrote a letter to the editor of the Boston Globe advocating that the comics page retain "Spiderman[47].") Individuals may also publish their own writing on personal web sites ("home pages" and "blogs").

Websites cannot be relied upon to archive their own material, because the websites may not exist in the future. For example, in the late 1990s thousands of articles and columns by leading writers were published at HotWired, a web property operated by Wired News. Wired News was eventually sold to Lycos, then to Conde Nast[38]. Numerous articles were lost during these transfers; those that are still available online are not at their original Internet location (`http://www.hotwired.com`), but are now housed underneath the `http://www.wired.com` domain. Many links *to*, *between* and even *within* the articles have been broken as a result.

One way to retrieve no longer extant web pages is through the use of the Internet "WayBack Machine," operated by the Internet Archive[3]. But here there are several problems:

- The Internet Archive is itself another organization (in this case a for-profit business) which may cease operation at some point in the future.
- The Archive's coverage is necessarily incomplete.
- The Internet Archive may not be accurate. (Fred Cohen has demonstrated that the content of "past" pages on the Internet Way Back machine can be manipulated from the future—a disturbing fact when one considers that the reports from WayBack machine have been entered into evidence in legal cases without challenge from opposing counsel[13].)
- The WayBack machine will not archive websites that are blocked with an appropriate *robots exclusion file* `robots.txt`. This was especially a problem for the "Journalspace" online journal, which was wiped out on January 2, 2009 due to an operator error and the lack of backups[43]. As it turns out, Journalspace

had a `robots.txt` file that prohibited archiving by services such as Internet Archive and Google.

Rather than hoping that another organization has managed to sweep up an individual's relevant web pages in a global cataloging of the Internet, it almost certainly makes more sense for archivists to go out and get the material themselves.

The Public Footprint may also contain information at social networking websites such as Facebook, MySpace and LinkedIn. These websites contains not just information that a person posted, but documentation of a person's social network—their "friends" and associates—as well as documentation of a person's preferences in the form of "recommendations" messages. Websites such as Flickr and Picassa hold photographs that a person may have uploaded. What a treasure for future historians trying to understand the life of an individual! What a quandary for an archivist, for these websites actively encourage originators to intermix the personal and the professional. Only through consultation with families and other interested parties will archivists be able to determine which "personal" information should be made immediately available, which information should be kept in closed collections until a suitable amount of time has passed, and what should be destroyed.

Finally, a person's public footprint might contain information that the person thinks is private but which is, in fact, public. It is notoriously difficult to audit security settings because they are complex and not generally apparent within today's user interfaces. As a result, it is common for computer users to make information publicly available when they do not intend to do so. Good and Krekelberg explored the Kazaa user interface and discovered that it was relatively easy for individuals to "share" their entire hard drive to a file sharing network when they intended to just share a few documents or folders[22]. Sometimes such inadvertent public sharing can have important political, social, or historical dimensions: in June 2008, Judge Alex Kozinski of the 9th US Circuit Court of Appeals was found to have sexually explicit photos and videos on his own personal website[31, 33][1]—relevant, as the Judge was himself overseeing an obscenity trial.

[1]Later the Judge defending himself saying that much of the material attributed to him by the Los Angeles Times had actually been posted by his son[25].

2.2 The Organizational Footprint

Although not strictly part of the "Internet" footprint, many organizations operate their own data services on which an originator could easily store information. For example, many businesses and organizations run their own web-based calendar and email services. These services may also cause problems for archivists because they can be hard to find and may not be readily interested in sharing their information—even when the originator or the originator's family strongly favor information sharing.

2.3 The Pseudonymous Footprint

Beyond the information that a person published under their own name, there is potentially a wealth of information that is publicly available but published under a different name or a non-standard email address—an electronic pseudonym.

There are may reasons why an individual might publish information to the public using a pseudonym:

- Information might be published under a different name in an attempt to preserve privacy.
- The individual might have a well-established pen name (for example, Charles Lutwidge Dodgson blogging as Lewis Caroll).
- The individual might be a fiction writer and be publishing the information online using the persona of a fictional character (for example, Dodgson blogging as the Queen of Hearts).
- The information might appear in an online forum where there is a community norm that prohibits publishing information under a "real name," or the online forum might assign pseudonyms as a matter of course.
- Another person might already be using the individual's name, forcing the originator to pick a different name.
- The individual might be a government or corporate official and be prohibited from posting under their own name for policy reasons. (For example, Whole Foods President John P. Mackey blogged under the pseudonym Rahobed, a play on his wife's name Deborah[35].)

Information that an originator publishes on the Internet in a manner that is freely available but is not directly linked to the person's name can be thought of as the individual's Pseudonymous Footprint. It is unlikely that all of

an originator's pseudonyms would be known in advance by an archivist: many people don't even remember all of the pseudonyms that they themselves use!

Pseudonyms have many characteristics that are sure to cause problems for future archivists:

- Although each pseudonym is typically used by a single person, this is not necessarily the case.
- Although some pseudonyms are long-lived, others may be created for a single purpose and then quickly discarded.
- Pseudonyms may be linguistically similar to the originator's name, similar to another person's name, or they may be unique.
- There is no central registry of pseudonyms.
- Some pseudonyms may be confined to a single online service, while others may be used between multiple services.
- The same pseudonym on different services may in fact be used by different people (e.g. while the user "rahobed" on Yahoo Finance bulletin was used by John P. Mackey, the blog `http://rahobed.blogspot.com/` actually belongs to one of the authors of this article.
- Pseudonyms that appear linked to email addresses (*e.g.* `rahobed@yahoo.com`) need not be: some online services allow *any* text string to be used as a username, and usernames that look like email addresses are not verified.

Automated tools may assist the researcher in attempting to determine if a pseudonym is or is not the originator[30]. In the case of photos, face recognition/matching software could be used.

2.4 The Anonymous Footprint

Anonymous works are fundamentally different from pseudonymous works. With pseudonymous messages there is at least a name ("Lewis Carroll") that the archivist can use to link a work to the true author. But for works that are truly anonymous, the only information that can link the work with the author is the content of the work itself.

Although the Internet originally had many outlets for anonymous speech, these systems received significant abuse as the Internet's popularity grew in the 1990s[26, 37]. As a result today's Internet has surprisingly few outlets for speech and messages that are truly anonymous.

2.5 The Private Footprint

Increasingly computer users are storing information on remote servers rather than on their own systems. Such services are sometimes called "grid," "cluster" or "cloud computing." Although these are online services, they are frequently used for private use. Individuals prefer them to using personally owned computer systems because of data durability (users don't need to back up their own data), and cost (most of the web-based services are free). Another advantage is that the systems make it relatively easy to collaborate with a small number of people.

Some examples of these services includes:

- Calender services (e.g. Google Calendar and Yahoo Calendar), which allows a person to have an online calendar.
- Online word processors and spreadsheets, such as Google Docs, and ThinkFree Boundless,
- Livejournal, a blogging service, which also allows for the creation of a private diary or a password-protected journal that is shared with a small number of people.
- Online banking and bill payment services. Whereas traditionally a person might have kept their own financial records, increasingly individuals are opting to receive "e-statements." Although e-statements could be sent by email, in practice the statements are not sent at all. Instead the bank or financial institution sends a message stating that the statement may be viewed on a website. Most users do not download a copy, but simply refer to the online version when they need to.

Access to online private services is typically protected with a username and a password. Most services allow users to register and email address; if a password is lost, a new password can be generated and sent to the address.

Also part of the private footprint are Internet services that do not appear as content at all—but which can be vital to understanding a person's approach to the online world. Two examples come to mind:

1. For example, Individuals can obtain domain name and populate the Domain Name System (DNS) database with a variety of types of information. Any attempt to capture Internet services which does not capture DNS is necessarily incomplete and may even be erroneous.But capturing only DNS is insufficient: there is necessarily a link between DNS names, IP

addresses, and geographical locations. Thus, in order to make sense of DNS information, it may be necessary to perform other operations such as geolocation[24] or cryptographic operations[16].

2. Much collaborative work that takes place on the Internet today is the collaborative creation of open source computer programs. These systems reside on servers such as SourceForge and Google Code, as well as on privately-managed CVS and Subversion servers. This code is generally not archived or indexed by existing search engines or web archiving projects, but may nevertheless have significant historical importance.

3 Finding the Footprint

As the previous section shows, simply mapping out the potential of a person's Internet Footprint is quite difficult. Actually finding it is more difficult still.

We have identified three approaches for finding an Internet Footprint: forensic analysis of an originator's computer system; search; and social network analysis.

3.1 Interviews with the Originator

Ideally, the originator or the originator's family will be able to provide a list of online services, complete with usernames and passwords, to enable the expeditious downloading and archiving of information stored on remote services. Such a list should also come with signed consent giving full authorization for the accounts to be used for the downloading of the information that they contain (see Section 5.1).

But even if the originator is alive and cooperating, it is unlikely that the originator will be able to provide a complete list of online information—most of us are simply unaware of all the various online services that we use on a daily basis. Finally, there is always the risk that the originator will have died without clearly documenting what online services were used. Even if the originator's family wishes to assist the archivist, they may be unable to do so.

Interviews may also be conducted with the originator's family and friends to see if they know of any online resources used by the originator.

3.2 Forensic Analysis

One of the most direct ways to identify an originator's Internet footprint is to conduct a forensic analysis of the originator's computers and other electronic devices.

Computer systems preserve many traces or remnants that are indicative of Internet activity:

- Web browsers maintain bookmarks and caches of web pages. Web pages may also be recovered from deleted files.
- Email messages are rich with references to online services in the form of emails containing links, notifications, password reset instructions.
- Address books may contain URLs and are frequently used to hold user names and passwords as well.
- Calendars may contain URLs and online information in their desktop calendars.
- Other references may be found in logfiles and even word processing documents.

Much of these references can be found by making a forensic copy of the originator's computer and all associated media (tapes, CD/DVDs, external drives etc), and then scanning the resulting disk images with a forensic feature extractor[19]. We have developed a primitive extractor called `bulk_extractor` which can produce a report of all email addresses and URLs found on an originator's hard drive. An example of the report of this program is shown in Figure 1.

Unfortunately, while some of an originator's account names, aliases, and pseudonyms may be present on the originator's machine, others may not be. The originator may have explicitly attempted to hide them, or may have accessed them exclusively from another machine, or they may have been used so long ago that references to the accounts have been overwritten.

The forensic analysis process should be completed with care not to alter or otherwise disturb the information on the originator's equipment. In general there are three key requirements which must be adhered to when conducting the analysis:

1. The entire storage space of the originator's computer and associated media should be captured, not merely the individual files. If possible, all attempts to copy data from the originator's computer should be done with a hardware write blocker in place between the computer and the storage media. This will ensure that data is not accidentally written *to* the originator's storage devices during the imaging process.

   Complete imaging of the originator's computer will establish the provenance of the captured material and address concerns of authenticity. These concerns are

```
Input file: /Users/simsong/M57 Jean.vmwarevm/Windows XP Clean-s001.vmdk
Starting page number: 0
Last processed page number: 90
Time: Fri Jan 16 11:59:27 2009

Top 10 email addresses:
=======================
jean@m57.biz: 1011
bob@m57.biz: 136
alex@m57.biz: 92
JEAN@M57.BIZ: 82
alison@m57.biz: 73
carol@m57.biz: 63
alison@M57.BIZ: 60
googlealerts-noreply@google.com: 49
inet@microsoft.com: 46
ca@digsigtrust.com: 40

Top 10 email domains:
=====================
m57.biz: 1487
M57.BIZ: 213
google.com: 84
netscape.com: 75
microsoft.com: 68
mozilla.org: 52
thawte.com: 51
digsigtrust.com: 46
hotmail.com: 35
aol.net: 31

Top 10 URLs:
====================
http://pics.ebaystatic.com/aw/pics/s.gif: 5056
http://www.microsoft.com/contentredirect.asp.: 1735
https://www.verisign.com/rpa: 673
http://www.mozilla.org/keymaster/gatekeeper/there.is.only.xul: 542
http://ocsp.verisign.com0: 526
http://: 430
http://support.microsoft.com: 424
http://pics.ebaystatic.com/aw/pics/paypal/logo_paypalPP_16x16.gif: 333
http://crl.verisign.com/ThawteTimestampingCA.crl0: 263
http://crl.verisign.com/tss-ca.crl0: 262
```

Figure 1: The first page of output from `bulk_extractor` program; the actual output runs more than 40 pages.

similar to those of legal authorities[2]. It may also result in data being preserved that would otherwise be lost—for example, residual data in deleted web browser cache files may contain important clues for uncovering pseudonyms used by the originator.

2. Data, once captured, should be "hashed," or cryptographically fingerprinted, with a strong algorithm such as SHA1 or SHA256. (MD5 is no longer sufficient as the algorithm has been compromised[49].) Even better, the image can be digitally signed and/or encrypted using a system such as the Advanced Forensic Format (AFF)[21].
3. In addition to a sector-by-sector copy of the storage media, it may be desirable to make a file-by-file copy. This will both assure that there are two copies of each file (one in the disk image and one in the copy), and will also decrease demands for the use of forensic tools. Also, in some cases, forensic tools may not be able to extract information from the disk images. (For example, in some cases it is not possible to easily reconstruct a multi-drive RAID or encrypted file system. In these cases it is easiest to use the host operating system to make a file-by-file copy.)

3.3 Search and Social Network Analysis

Another way to locate the originator's Internet footprint is by searching for it. Two kinds of search are possible. First, the archivist could simply search for the originator's name (or aliases) on Internet search systems such as Google and Yahoo. Second, the archivist could go specifically to websites such as Facebook, MySpace and Flickr, and conduct searches there.

Search is complicated by the fact that many people share the same name. Bekkerman and McCallum note that a search for the name "David Mulford" on Google correctly retrieves information about a US Ambassador to India, "two business managers, a musician, a student, a scientist, and a few others"—all people who share the same name[8]. Which David Mulford is the "right" David Mulford depends on which one the context of the search.

Sometimes it is difficult to determine if two seemly different individuals are in fact the same person. Consider again the search for "David Mulford:"

> "It is sometimes quite difficult to determine if a page is about a particular person or not. In case of Ambassador David Mulford, much of the information that can be found at first may seem to be unrelated: one site states that in the late 1950s David attended Lawrence University and was a member of its athletic team; other sites mention his work at different positions in governmental departments and commercial structures, including Chairman International of Credit Suisse First Boston (CSFB) in London; a few sites (mostly in Spanish) relate his name to a financial scandal in Argentina. It is a diffcult challenge to automatically determine whether all of these sites discuss the same person."[8]

The archivist can also try to find an originator's Internet footprint by searching the websites belonging to the originator's known friends and relations and looking for links. In some cases it may be appropriate to directly email individuals in the originator's address book or social network to see if they have information that they wish to share with the archivist.

Once references are found, it might be useful to sort these references into a variety of categories. We suggest three:

Provable References Known references could be indicated by the presence of a username/password combination which maps directly to a specific website and can be validated by testing to see if the account can still be accessed.

Reliable References A reliable reference could be indicated by the presence of an alias and URL/cookie combination but does not include a password, preventing the researcher from actually testing the account.

Passing References A passing reference could be indicated by the presence of a URL or cookie which points to a social networking site or internet e-mail site. The difference here is that there is only one indicator of reference to a website which could hold historically interesting material.

3.4 Unexpected Complications

3.4.1 Comments, Tracebacks, and Diggs

Now, think back to the BBC and CNN news sites. Although these services *seem* to be anonymous publication services, increasingly these services are places where an originator may leave an Internet footprint. BBC's website allows users to create a membership, "Sign In" and leave comments on every story. Comments are displayed with the user's "member name" which is *unique*. an originator might use his or her real name as a member name. Alternatively, the originator might use a pseudonym (or multiple pseudonyms) which might or might not be similar to the originator's real name. A future biographer trying to build a picture of the originator might be very interested in the comments that the person thought to leave on the BBC website—putting those comments in context requires not just archiving them, but archiving the original story and the other comments as well.

CNN also allows readers to post a comment (or "Sound Off," to use CNN's term). But CNN also allows users to share articles on services such as Mixx, Digg, Facebook, del.icio.us, reddit, StumbleUpon, and MySpace. "Sharing" means that a reference to the article, and the user's comments about the article, are cross-posted to another web-based service.

3.4.2 "Report as Offensive" and Edit Wars

Another complication is that user contributions may be removed by other users. Web sites have given users this power to manage the torrents of spam and inappropriate comments that many high-profile websites receive. For example, the BBC website allows users to "Complain about this comment" (Figure 3), and Craigslist allows comments to be flagged as "miscategorized," "prohibited," or "spam/overpost" (Figure 2). Many websites will automatically remove user-generated comment that is flagged by more than a certain number of people.

On Wikipedia it is even easier to change an originator's words—they can simply be edited by other Wikipedia users. This is particularly problematic when people are contributing to articles that are controversial. Imagine a noted author or historian locked in a bitter "edit war" with some other Wikipedia user, with each editing and re-editing the works of the other. Then the noted historian dies. With no one left to defend the historian's intellectual space, the pages get rewritten or even marked for deletion

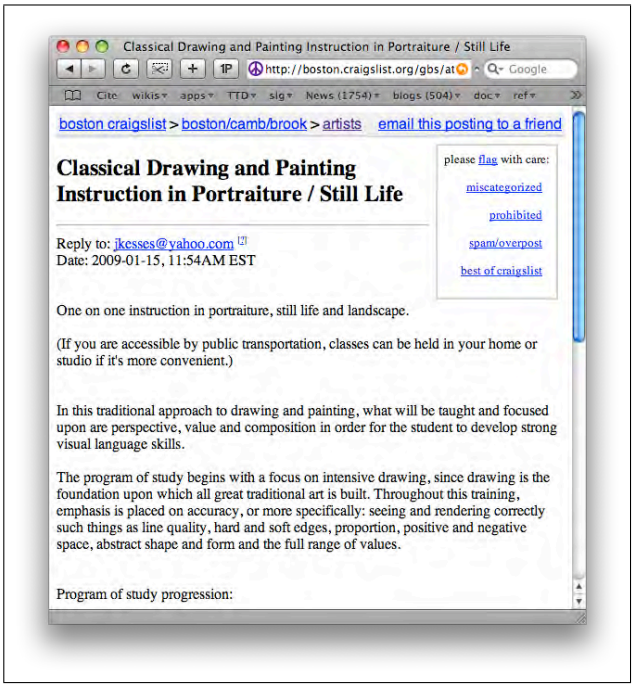

Figure 2: Postings to Craiglist may one day provide fascinating contemporanious documents of the career of writers or artists.

and are eventually removed from the system. From the point of view of Wikipedia policy this is the correct outcome, as a Wikipedia article is supposed to represent a consensus truth that can be verified from external sources and for which the author has no vested interest[20].

3.4.3 Privacy Enhancing Technologies

The originator may have employed various privacy enhancing technologies (PETs) such as encryption or anonymity services during their lifetime. Such services, unfortunately, may also prevent the analysis of their computer systems by archivists after the originator's death. This can be a problem even if the analysis is performed with the full consent of the originator's family.

For example, data may be encrypted, either on the originator's home computer system or on remote servers. In recent years high-quality encryption has been built into consumer operating system (for example, Apple's FileVault). There are also a small number of Internet service

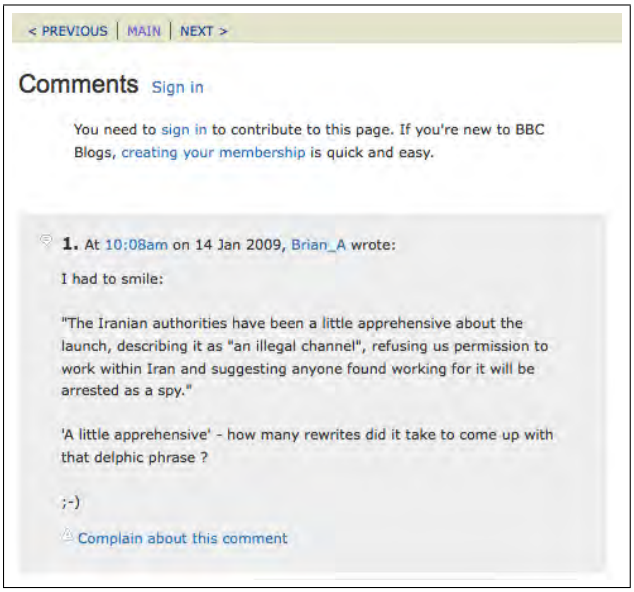

Figure 3: The BBC website allows users to complain about comments left from other users.

providers that offer to store information in an encrypted form so that not even the provider can access it (for example, HushMail offers encryption of email, while Iron Mountain Digital Services offers encryption of backups.)

Encryption may be subverted through the analysis of the originator's own computer systems, as sometimes people store passwords and encryption keys for remote systems on their local computers. Programs such as AccessData's Forensic Tool Kit and Password Recovery Tool Kit can work together to scan a hard drive for proper names, use this information to try to forcibly decrypt, or "crack," the encrypted data. The company's Distributed Network Attack can run the attack simultaneously on hundreds of computers to dramatically increase speed.

Crack today or crack tomorrow? Archivists have an interesting dilemma when attempting to decrypt encrypted data. In most cases it becomes easier to forcibly decrypt encrypted data as each year computers get faster and new techniques are discovered for cracking. On the other hand, a lessor-known encryption technique may conceivably become more difficult to decrypt with the passage of time as the number of people familiar with the specific technique dwindles. It is possible that a weak but obscure algorithm that is crackable today will not be readily crackable in the future without significant re-investment in research as the specific knowledge of the vulnerability is lost.

3.4.4 Uncooperative Service Providers

There is an old story of an assistant at MIT who worked for a famous professor in one of the physical science departments. One day the professor died after a long illness. Shortly thereafter, the assistant received a phone call from the Institute Archivist who wanted to stop by and evaluate the professor's papers. The assistant said that she had been expecting the archivist and had already "cleaned them up" in anticipation of the visit. When the archivist arrived the extent of the cleaning became evident: the assistant had thrown out the professor's scratch pads, his doodles, a box of business receipts, and so on, and prepared for the archivist a neat folder showing all of the professor's speeches, published articles, and honors. The archivist was devastated.

Although many archivists know that they may need to act with haste in order to preserve the physical papers of the deceased, this story of the archivist and the assistant is in danger of playing out with great frequency in tomorrow's cloud-based world of electronic records.

For example, photo sharing websites such as AOL Pictures have deleted uploaded pictures that are not viewed after 60 days, or when the owner of the account fails to log in after 90 days. Some services delete photos when monthly fees are no longer paid[10]. Archivists would need to move fast to rescue an originator's photos stored on such a service.

Facebook's policy is to place the profile of members who die into a *Memorial State.* "In Memorial State, the account is given stronger privacy settings (only friends can see the profile), the person is removed from any groups and the status is taken away. This policy is the same across the board. If the family would rather the profile be taken down, we will do so," stated Malorie Lucich, a spokesperson for the company[34].

But Facebook's only changes the account to memory state if someone brings to Facebook's attention that a member has died. Meanwhile, an article at the University of Georgia's newspaper details how parents of deceased students have taken over their Facebook accounts, using

the service as a means for memorializing their children and getting to know their children's friends[27].

4 Archiving the Footprint

Information must be archived once it is discovered. Archiving involves two distinction processes: getting the content, and saving the content.

4.1 Getting the content

Once the references have been cataloged, the archivist must then begin the task of extracting content from the Internet and saving it in an archival form. The archivist can manually log into the remote websites to access the information or, more likely, run some kind of modified web crawler (e.g. [39]) to do the work.

For historic purposes it will almost always be desirable to store the original web page. However, since many web pages are likely to contain extraneous information (*e.g.* advertisements and navigation elements), it may also be desirable to automatically extract the relevant portions of a web using a "wrapper" or information extractor. Generally though these tools are hand written to suit a specific web site and do not scale or transfer well from page to page. Fortunately, tools have been proposed to better address the issues associated with wrapper development, including W4F (World Wide Web Wrapper Factory)[44], Rapier (Robust Automated Production of Information Extraction)[12] or NoDoSE (Northwestern Document Structure Extractor)[6].

HTML-aware tools, like W4F, typically provide a higher degree of automation; however, the consistent use of HTML tags on target pages is required. Tools which are based on Natural Language Processing (NLP), such as RAPIER, can be classified as semi-automatic because though the wrapper is generated automatically a user needs to provide examples to guide the it. It is up to the researcher to choose (or develop) the appropriate tool. (A comprehensive list of information extraction approaches can be found in [46].)

Reliable References will require a more hands on approach. This category will require the archivist to manually navigate to the website and identify whether or not it is historically interesting. If it is deemed so then the tools discussed in the previous category may certainly be used to extract appropriate content ensuring that appropriate steps are taken to maintain an original copy and integrity assurances.

The third category, Passing References, will require significant time and effort on the part of the historian and it is anticipated that the level of automation will decrease. Since the historian is provide little information on which to go on exhaustive manual searches of both local and deep/hidden content will be required. For public content the use of traditional search engines, like Google and Yahoo, and Webcrawlers, like Webcrawler.com and DataRover could be utilized. As local search engines index mostly based on hyperlinks which include location information they typically exclude high quality "local" content available in the Deep Web[40]. Deep Web crawling may be accomplished through the use of tools such as Deep Web Crawler and LocalDeepBot. Additionally, Hidden Web Agents may be used as well. These agents can search and collect information on pages outside the Publically Indexable Web (PIW)[32].

4.2 Saving the content

While there are many different ways to archive web content, each has significant technical problems.

There are several fundamental problems in making an archival copy of a web page:

- Because web pages can appear differently on different computers, it is not clear what should be archived—a picture of the web page, or the HTML code of a web page?
- Web sites such as Facebook and LiveJournal may show web pages differently depending on who is logged in. Should the web page be archived as it appear to the author, to a person in the author's circle of "friends," to an un-friended registered user, or as it appear if no one is logged in?
- Alternatively, web sites may display pages differently at different times of day, or change their "theme" to take into account current events. If there are significant time-dependent changes, should multiple copies be archived?

Once the archivist decides *what* should be archived, the next question to answer is *how* should it be archived.

The naïve approach for archiving web content is to print it. While archivists generally frown on this approach, because all it does is exchange one set of problems for another.

Instead of printing to paper, the web page could be

"printed" to a bitmap file (e.g. a TIFF or PNG). Such files produce an exact copy of what was seen on the screen—at least for one specific web browser—but they cannot be readily searched unless they are OCRed. Such scans do, however, meet the legal requirements for admission to the US courts[17].

Another approach is to "print" the web content to Adobe Acrobat (PDF) format. But PDF is an evolving standard: PDF documents created today may look differently in 10 years with a different Acrobat reader. Acrobat has specifically had problems with documents that had embedded bitmap fonts (especially documents created by versions of LaTeX in the 1980s and 1990s) and documents authored in languages other than English which did not have embedded fonts[45].

5 Legal Issues

There are primarily two legal issues that could arise during the conduct of the collection of Internet works being proposed in this paper: violations of copyright law, and violations of computer crime statutes such as the US Computer Fraud and Abuse Act, or the UK Computer Misuse Act. There are also a number of ethical issues that might arise as well.

5.1 Copyright and Terms of Use

Copyright law, at least in the United States, is generally quite receptive to archives made for scholarly purposes, especially when the archiving is done for non-commercial purpose and in such a way that the value of the original copyrighted work is not compromised. In such a case, copies are typically allowed under the "Fair Use" doctrine (17 U.S.C. §106); similar Fair Use is allowed under other copyright regimes as well.

Despite Fair Use, many web publishers and online services are generally not receptive to having their content scraped, spidered, or otherwise archived. For example, Facebook's *Terms of Use* (Figure 4)clearly prohibits archiving an originators Facebook postings by anyone other than the person herself; whether or not this permission would apply to the person's estate or an archivist acting on behalf of the person or estate is unclear. However, the policy is very clear that Facebook would not permit an archivist or historian to archive and then display messages that others had posted on the originator's Facebook "wall," or messages that the person had received, or how the originator's Facebook presence existed in the context of other Facebook profiles.

> "No Site Content may be modified, copied, distributed, framed, reproduced, republished, downloaded, scraped, displayed, posted, transmitted, or sold in any form or by any means, in whole or in part, without the Company's prior written permission, except that the foregoing does not apply to your own User Content (as defined below) that you legally post on the Site. Provided that you are eligible for use of the Site, you are granted a limited license to access and use the Site and the Site Content and to download or print a copy of any portion of the Site Content to which you have properly gained access solely for your personal, non-commercial use, provided that you keep all copyright or other proprietary notices intact. Except for your own User Content, you may not upload or republish Site Content on any Internet, Intranet or Extranet site or incorporate the information in any other database or compilation, and any other use of the Site Content is strictly prohibited[18]."

Figure 4: This section of Facebook's *Terms of Use* would seem to prohibit the archiving of a person's Facebook profile for historical purposes.

5.2 Computer Crime

Even if an archivist decided that it is legally permissible to archive the content that an originator may have stored in the "Internet cloud," the way that the archivist goes about performing this function may expose the archivist to criminal charges.

For example, although it may be possible to scan an originator's hard drive for the username and password to an online service, actually using that username and password may put the archivist in violation of computer crime statutes such as the US Computer Fraud and Abuse Act (CFAA) (18 USC §1030). Such violations may be direct, as the CFA prohibits "unauthorized access" to computers involved in interstate commerce. But violations may also be indirect, the result of violating a website's "Terms of Service" under a growing interperation of the CFAA which holds individuals criminally liable for using a website in a manner other than that which was envisioned by its the website's owner[14].

5.3 Ethical Issues

Computer systems have the potential to record more information, retain it for a longer period of time, and make it available to more individuals than is possible with paper works. More than ever, every effort should be made to clearly differentiate between what is public and what is private information. This is especially the case when collecting from online information systems, since there is the chance that the information collected may belong to another person (in the case of a mistaken identity), or may involve other people (in the case of a social network website).

The problem of mistaken identity is especially problematic for online data collection. There is little chance when going through a person's office that the archivist will accidently pick up and catalog a diary belonging to a person who has the same name but who lives in another country—but this is exactly what can happen when downloading a originator's online diary.

6 Conclusion

It is no longer sufficient to simply analyze local computers and associated media when attempting to catalog a persons life works. Ever increasingly communication, personal documents and published works are migrating to the web space. Social Networking sites contain photos, videos and personal communication. Blog sites contain personal ramblings and commentaries; named and anonymous. E-mail and chat as well as personal videos are also migrating to the web. The archivist of the present must be technically savvy and be able to use the myriad of forensic analysis, web searching and cataloging tools in order to be efficient and create a complete set of works.

Many of the approaches discussed in this paper need not be confined to the archivist profession. Individuals can apply these approaches on themselves to determine the extent of their own digital shadow. These approaches may also be useful in civil litigation for e-discovery, and even in law enforcement.

6.1 Acknowledgements

Our thanks Jeremy Leighton John at the Digital Lives research project for suggesting that we explore this relevant and interesting topic and providing valuable feedback on this paper.

References

[1] Memory of the world: Safeguarding the documentary heritage: A guide to standards, recommended practices and reference literature related to the preservation of documents of all kinds., July 15 2003. http://www.unesco.org/webworld/mdm/administ/en/guide/guidetoc.htm.

[2] Adapting existing technologies for digitally archiving personal lives. In *iPRES 2008: The Fifth International Conference on Preservation of Digital Objects*. The British Library, September 2008. http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf.

[3] Internet archive wayback machine, 2008. http://web.archive.org.

[4] Salon magazine: Breaking news, opinion, politics, entertainment, sports and culture, 2009. http://www.salon.com/.

[5] Slate magazine, 2009. http://www.slate.com.

[6] Brad Adelberg. Nodose—a tool for semi-automatically extracting structured and semistructured data from text documents. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 283–294. ACM, New York, NY, USA, 1998. ISBN 0-89791-995-5.

[7] Ziv Bar-Yossef, Andrei Z. Broder, Ravi Kumar, and Andrew Tomkins. Sic transit gloria telae: towards an understanding of the web's decay. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 328–337. ACM, New York, NY, USA, 2004. ISBN 1-58113-844-X.

[8] Ron Bekkerman and Andrew McCallum. Disambiguating web appearances of people in a social network. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 463–470. ACM, New York, NY, USA, 2005. ISBN 1-59593-046-9.

[9] Susan L. Bryant, Andrea Forte, and Amy Bruckman. Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 1–10. ACM, New York, NY, USA, 2005. ISBN 1-59593-223-2.

[10] William M. Bulkeley. Failure to log on, buy prints can lead to loss of pictures; wife 'on the verge of tears'. *The Wall Street Journal*, February 1 2006. http://www.phanfare.com/press/wsj_bulkeley.pdf.

[11] Moira Burke and Robert Kraut. Taking up the mop: identifying future wikipedia administrators. In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, pages 3441–3446. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-012-X.

[12] Mary Elaine Califf and Raymond J. Mooney. Bottom-up relational learning of pattern matching rules for information extraction. *J. Mach. Learn. Res.*, 4:177–210, 2003. ISSN 1533-7928.

[13] Fred Cohen. Risks of believing what you see on the wayback machine (archive.org). *RISKS Digest*, 25, January 7 2008. http://seclists.org/risks/2008/q1/0000.html.

[14] Susan Crawford. The computer fraud and abuse act, May 19 2008. http://scrawford.net/blog/the-computer-fraud-and-abuse-act/1172/.

[15] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008. ISSN 0360-0300.

[16] Mark Delany. Domain-based email authentication using public-keys advertised in the DNS (domainkeys), August 2004. INTERNET DRAFT.

[17] John P. Elwood. Admissibility in federal court of electronic copies of personnel records, May 30 2008. http://www.usdoj.gov/olc/2008/electronic-personnel-records.pdf.

[18] Facebook. Terms of use, September 28 2008. http://www.facebook.com/terms.php.

[19] Simson Garfinkel. Forensic feature extraction and cross-drive analysis. In *Proceedings of the 6th Annual Digital Forensic Research Workshop (DFRWS)*. Lafayette, Indiana, August 2006. http://www.dfrws.org/2006/proceedings/10-Garfinkel.pdf.

[20] Simson L. Garfinkel. Wikipedia and the meaning of truth. *Technology Review*, November/December 2008. https://www.technologyreview.com/web/21558/.

[21] Simson L. Garfinkel. Providing cryptographic security and evidentiary chain-of-custody with the advanced forensic format, library, and tools. *The International Journal of Digital Crime and Forensics*, 1, January–March 2009.

[22] Nathaniel S. Good and Aaron Krekelberg. Usability and privacy: a study of Kazaa P2P file-sharing. In *Proceedings of the conference on Human factors in computing systems*, pages 137–144. ACM Press, 2003. ISBN 1-58113-630-7.

[23] Google. Google calendar help, 2009. http://www.google.com/support/calendar/.

[24] Saikat Guha and Paul Francis. In *Privacy Enhancing Technologies*, pages 153–166. Springer, 2007. http://www.cs.cornell.edu/people/francis/pet07-idtrail-cameraready.pdf.

[25] Karen Gullo. Panel of five to probe judge's sexual web postings. *Bloomberg*, June 17 2008. http://www.bloomberg.com/apps/news?pid=newsarchive\\&sid=ahD1O6qXYiGc.

[26] Sabine Helmers. A brief history of anon.penet.fi - the legendary anonymous remailer, September 1997. http://www.december.com/cmc/mag/1997/sep/helmers.html.

[27] Brian Hughes. Facebook for the great beyond, July 11 2007. http://media.www.redandblack.com/media/storage/paper871/news/2007/11/07/Variety/Facebook.For.The.Great.Beyond-3083145.shtml.

[28] Adam Jatowt, Yukiko Kawai, and Katsumi Tanaka. Detecting age of page content. In *WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management*, pages 137–144. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-829-9.

[29] Jeremy Leighton John. Adapting existing technologies for digitally archiving personal lives. In *iPres 2008*, 2008. http://www.bl.uk/ipres2008/programme.html.

[30] Patrick Juola. Authorship attribution. *Found. Trends Inf. Retr.*, 1(3):233–334, 2006. ISSN 1554-0669.

[31] Judge alex kozinski calls for probe into his porn postings. *Los Angeles Times*, June 13 2008.

[32] Juliano Palmieri Lage, Altigran S. da Silva, Paulo B. Golgher, and Alberto H. F. Laender. Automatic generation of agents for collecting hidden web pages for data extraction. *Data Knowl. Eng.*, 49(2):177–196, 2004. ISSN 0169-023X.

[33] Lawrence Lessig. The Kozinski mess, June 12 2008. http://www.lessig.org/blog/2008/06/the_kozinski_mess.html.

[34] Malorie Lucich. Re: face book pages of dead people, January 16 2009. personal communication.

[35] Andrew Martin. Whole foods executive used alias. *The New York Times*, July 12 2007. http://www.nytimes.com/2007/07/12/business/12foods.html.

[36] Frank McCown, Norou Diawara, and Michael L. Nelson. Factors affecting website reconstruction from the web infrastructure. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 39–48. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-644-8.

[37] Declan McCullagh. Finish line, September 1996. `http://w2.eff.org/Misc/Publications/Declan_McCullagh/hw.finnish.line.090696.article`.

[38] Elinor Mills. Conde nast to buy wired news, July 11 2006. `http://news.cnet.com/Conde-Nast-to-buy-Wired-News/2100-1030_3-6093028.html`.

[39] José E. Moreira, Maged M. Michael, Dilma Da Silva, Doron Shiloach, Parijat Dube, and Li Zhang. Scalability of the nutch search engine. In *ICS '07: Proceedings of the 21st annual international conference on Supercomputing*, pages 3–12. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-768-1.

[40] Dheerendranath Mundluru and Xiongwu Xia. Experiences in crawling deep web in the context of local search. In *GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 35–42. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-253-5.

[41] Maureen Pennock and Brian Kelly. Archiving web site resources: a records management view. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 987–988. ACM, New York, NY, USA, 2006. ISBN 1-59593-323-9.

[42] Herman Chung-Hwa Rao, Yih-Farn Chen, and Ming-Feng Chen. A proxy-based personal web archiving service. *SIGOPS Oper. Syst. Rev.*, 35(1):61–72, 2001. ISSN 0163-5980.

[43] Craig Richmond. Why mirroring is not a backup solution. January 2 2009. `http://hardware.slashdot.org/article.pl?sid=09%2F01%2F02%2F1546214`.

[44] Arnaud Sahuguet and Fabien Azavant. Building light-weight wrappers for legacy web data-sources using w4f. In *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*, pages 738–741. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999. ISBN 1-55860-615-7.

[45] Adobe Systems. Adobe acrobat 4.0 for macintosh readme, March 15 1999.

[46] Jordi Turmo, Alicia Ageno, and Neus Català. Adaptive information extraction. *ACM Comput. Surv.*, 38(2):4, 2006. ISSN 0360-0300.

[47] John Updike. Cut the unfunny comics, not "spiderman". *The Boston Globe*, October 27 1994.

[48] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. ACM, New York, NY, USA, 2004. ISBN 1-58113-702-8.

[49] Xiaoyun Wang and Hongbo Yu. How to break md5 and other hash functions. In Ronald Cramer, editor, *EUROCRYPT*, volume 3494 of *Lecture Notes in Computer Science*, pages 19–35. Springer, 2005. ISBN 3-540-25910-4.

[50] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003. ISSN 0360-0300.